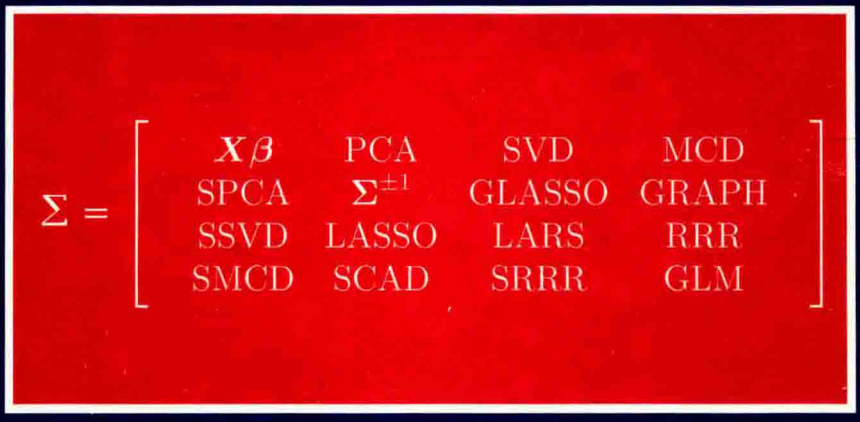
Wiley Series in Probability and Statistics

High-Dimensional Covariance Estimation



Mohsen Pourahmadi

WILEY

HIGH-DIMENSIONAL COVARIANCE ESTIMATION

MOHSEN POURAHMADI
Texas A & M University

WILEY

Published by John Wiley & Sons, Inc., Hoboken, New Jersey.
Published simultaneously in Canada.

Library of Congress Cataloging-in-Publication Data:

Pourahmadi, Mohsen.
Modern methods to covariance estimation / Mohsen Pourahmadi, Department of Statistics, Texas A&M University, College Station, TX.
pages cm
Includes bibliographical references and index.
ISBN 978-1-118-03429-3 (hardback)
1. Analysis of covariance. 2. Multivariate analysis. I. Title.
QA279.P68 2013
519.5′38–dc23

2013000326

Printed in the United States of America.

10 9 8 7 6 5 4 3 2 1

HIGH-DIMENSIONAL COVARIANCE ESTIMATION

WILEY SERIES IN PROBABILITY AND STATISTICS

Established by WALTER A. SHEWHART and SAMUEL S. WILKS

A complete list of the titles in this series appears at the end of this volume.

PREFACE

The aim of this book is to bring together and present some of the most important recent ideas and methods in high-dimensional covariance estimation. It provides computationally feasible methods and their conceptual underpinnings for sparse estimation of large covariance matrices. The major unifying theme is to reduce sparse covariance estimation to that of estimating suitable regression models using penalized least squares. The framework has the great advantage of reducing the unintuitive and challenging task of covariance estimation to that of modeling a sequence of regressions. The book is intended to serve the needs of researchers and graduate students in statistics and various areas of science, engineering, economics and finance. Coverage is at an intermediate level, familiarity with the basics of regression analysis, multivariate analysis, and matrix algebra is expected.

A covariance matrix, the simplest summary measure of dependence of several variables, plays prominent roles in almost every aspect of multivariate data analysis. In the last two decades due to technological advancements and availability of high-dimensional data in areas like microarray, e-commerce, information retrieval, fMRI, business, and economy, there has been a growing interest and great progress in developing computationally fast methods that can handle data with as many as thousand variables collected from only a few subjects. This situation is certainly not suited for the classical multivariate statistics, but rather calls for a sort of "fast and sparse multivariate methodology."

The two major obstacles in modeling covariance matrices are high-dimensionality (HD) and positive-definiteness (PD). The HD problem is familiar from regression analysis with a large number of covariates where the penalized least squares with the Lasso penalty is commonly used to obtain computationally feasible solutions. However, the PD problem is germane to covariances where one hopes to remove it by

infusing regression-based ideas into principal component analysis (PCA), Cholesky decomposition, and Gaussian graphical models (inverse covariance matrices), etc.

The primary focus of current research in high-dimensional data analysis and hence covariance estimation has been on developing feasible algorithms to compute the estimators. There has been less focus on inference and the effort is mostly devoted to establishing consistency of estimators when both the sample size and the number of variables go to infinity in certain manners depending on the nature of sparsity of the model and the data. At present, there appears to be a sort of disconnection between the theory and practice where further research is hoped to bridge the gap. Our coverage follows mostly the recent pattern of research in the HD data literature by focusing more on the algorithmic aspects of the high-dimensional covariance estimation. This is a rapidly growing area of statistics and machine learning, less than a decade old, but has seen tremendous growth in such a short time. Deciding what to include in the first book of its kind is not easy as one does not have the luxury of choosing results that have passed the test of time. My selection of topics has been guided by the promise of lasting merit of some of the existing and freshly minted results, and personal preferences.

The book is divided into two parts. Part I, consisting the first three chapters, deals with the more basic concepts and results on linear regression models, high-dimensional data, regularization, and various models/estimation methods for covariance matrices. Chapter 1 provides an overview of various regression-based methods for covariance estimation, Chapter 2 introduces several examples of high-dimensional data and illustrates the poor performance of the sample covariance matrix and the need for its regularization. A fairly comprehensive review of mathematical and statistical properties of the covariance matrices along with classical covariance estimation results is provided in Chapter 3. Part II is concerned with the modern high-dimensional covariance estimation. It covers shrinkage estimation of covariance matrices, sparse PCA, Gaussian graphical models, and penalized likelihood estimation of inverse covariance matrices. Chapter 6 deals with banding, tapering, and thresholding of the sample covariance matrix or its componentwise penalization. The focus of Chapter 7 is on applications of covariance estimation and singular value decomposition (SVD), to multivariate regression models for high-dimensional data.

The genesis of the book can be traced to teaching a topic course on covariance estimation in the Department of Statistics at the University of Chicago, during a sabbatical in 2001–2002 academic year. I have had the benefits of discussing various topics and issues with many colleagues and students including Anindya Bahdra, Lianfu Chen, Michael Daniels, Nader Ebrahimi, Tanya Garcia, Shuva Gupta, Jianhua Huang, Priya Kohli, Soumen Lahiri, Mehdi Madoliat, Ranye Sun, Adam Rothman, Wei Biao Wu, Dale Zimmerman, and Joel Zinn. Financial support from the NSF in the last decade has contributed greatly to the book project. The editorial staff at John Wiley & Sons and Steve Quigley were generous with their assistance and timely reminders.

MOHSEN POURAHMADI

College Station, Texas
April, 2013

CONTENTS

PREFACE ix

I MOTIVATION AND THE BASICS

1 INTRODUCTION 3

1.1 Least Squares and Regularized Regression / 4
1.2 Lasso: Survival of the Bigger / 6
1.3 Thresholding the Sample Covariance Matrix / 9
1.4 Sparse PCA and Regression / 10
1.5 Graphical Models: Nodewise Regression / 13
1.6 Cholesky Decomposition and Regression / 13
1.7 The Bigger Picture: Latent Factor Models / 15
1.8 Further Reading / 17

2 DATA, SPARSITY, AND REGULARIZATION 21

2.1 Data Matrix: Examples / 22
2.2 Shrinking the Sample Covariance Matrix / 26
2.3 Distribution of the Sample Eigenvalues / 29
2.4 Regularizing Covariances Like a Mean / 30
2.5 The Lasso Regression / 32
2.6 Lasso: Variable Selection and Prediction / 36

2.7 Lasso: Degrees of Freedom and BIC / 37
2.8 Some Alternatives to the Lasso Penalty / 38

3 COVARIANCE MATRICES 45

3.1 Definition and Basic Properties / 45
3.2 The Spectral Decomposition / 49
3.3 Structured Covariance Matrices / 53
3.4 Functions of a Covariance Matrix / 56
3.5 PCA: The Maximum Variance Property / 61
3.6 Modified Cholesky Decomposition / 63
3.7 Latent Factor Models / 67
3.8 GLM for Covariance Matrices / 73
3.9 GLM via the Cholesky Decomposition / 76
3.10 GLM for Incomplete Longitudinal Data / 79
3.10.1 The Incoherency Problem in Incomplete Longitudinal Data / 79
3.10.2 The Incomplete Data and The EM Algorithm / 81
3.11 A Data Example: Fruit Fly Mortality Rate / 84
3.12 Simulating Random Correlation Matrices / 89
3.13 Bayesian Analysis of Covariance Matrices / 91

II COVARIANCE ESTIMATION: REGULARIZATION

4 REGULARIZING THE EIGENSTRUCTURE 99

4.1 Shrinking the Eigenvalues / 100
4.2 Regularizing The Eigenvectors / 105
4.3 A Duality between PCA and SVD / 107
4.4 Implementing Sparse PCA: A Data Example / 110
4.5 Sparse Singular Value Decomposition (SSVD) / 112
4.6 Consistency of PCA / 114
4.7 Principal Subspace Estimation / 118
4.8 Further Reading / 119

5 SPARSE GAUSSIAN GRAPHICAL MODELS 121

5.1 Covariance Selection Models: Two Examples / 122
5.2 Regression Interpretation of Entries of Σ^{-1} / 124
5.3 Penalized Likelihood and Graphical Lasso / 126

5.4 Penalized Quasi-Likelihood Formulation / 131
5.5 Penalizing the Cholesky Factor / 132
5.6 Consistency and Sparsistency / 136
5.7 Joint Graphical Models / 137
5.8 Further Reading / 139

6 BANDING, TAPERING, AND THRESHOLDING **141**

6.1 Banding the Sample Covariance Matrix / 142
6.2 Tapering the Sample Covariance Matrix / 144
6.3 Thresholding the Sample Covariance Matrix / 145
6.4 Low-Rank Plus Sparse Covariance Matrices / 149
6.5 Further Reading / 150

7 MULTIVARIATE REGRESSION: ACCOUNTING FOR CORRELATION **153**

7.1 Multivariate Regression and LS Estimators / 154
7.2 Reduced Rank Regressions (RRR) / 156
7.3 Regularized Estimation of B / 158
7.4 Joint Regularization of (B, Ω) / 160
7.5 Implementing MRCE: Data Examples / 163
7.5.1 Intraday Electricity Prices / 163
7.5.2 Predicting Asset Returns / 165
7.6 Further Reading / 167

BIBLIOGRAPHY 171
INDEX 181

PART I

MOTIVATION AND THE BASICS

CHAPTER 1

INTRODUCTION

Is it possible to estimate a covariance matrix using the regression methodology? If so, then one may bring the vast machinery of regression analysis (regularized estimation, parametric and nonparametric methods, Bayesian analysis, ...) developed in the last two centuries to the service of covariance modeling.

In this chapter, through several examples, we show that sparse estimation of high-dimensional covariance matrices can be reduced to solving a series of regularized regression problems. The examples include sparse principal component analysis (PCA), Gaussian graphical models, and the modified Cholesky decomposition of covariance matrices. The roles of sparsity, the least absolute shrinkage and smooth operator (Lasso) and particularly the soft-thresholding operator in estimating the parameters of linear regression models with a large number of predictors and large covariance matrices are reviewed briefly.

Nowadays, high-dimensional data are collected routinely in genomics, biomedical imaging, functional magnetic resonance imaging (fMRI), tomography, and finance. Let $\boldsymbol{X}$ be an $n \times p$ data matrix where n is the sample size and p is the number of variables. By the *high-dimensional data* usually it is meant that p is bigger than n. Analysis of high-dimensional data often poses challenges which calls for new statistical methodologies and theories (Donoho, 2000). For example, least-squares fitting of linear models and classical multivariate statistical methods cannot handle high-dimensional $\boldsymbol{X}$ since both rely on the inverse of $\boldsymbol{X}'\boldsymbol{X}$ which could be singular or not well-conditioned. It should be noted that increasing n and p each has very different and opposite effects on the statistical results. In general, the focus of multivariate analysis is to make statistical inference about the dependence among variables so

High-Dimensional Covariance Estimation, First Edition. Mohsen Pourahmadi.

that increasing n has the effect of improving the precision and certainty of inference, whereas increasing p has the opposite effect of reducing the precision and certainty. Therefore the level of detail that can be inferred about correlations among variables improves with increasing n but it deteriorates with increasing p.

The dimension reduction and variable selection are of fundamental importance for high-dimensional data analysis. The *sparsity principle* which assumes that only a small number of predictors contribute to the response is frequently adopted as the guiding light in the analysis. Armed with the sparsity principle, a large number of estimation approaches are available to estimate sparse models and select the significant variables simultaneously. The Lasso method introduced by Tibshirani (1996) is one of the most prominent and popular estimation methods for the high-dimensional linear regression models.

Quantifying the interplay between high-dimensionality and sparsity is important in the modern data analysis environment. In the classic setup, usually p is fixed and n grows so that

$$\frac{p}{n} \ll 1.$$

However, in the modern high-dimensional setup where both p and n can grow, estimating accurately a vector β with p parameters is a real challenge. By invoking the *sparsity principle* one assumes that only a small number, say $s(\beta)$, of the entries of β is nonzero and then proceeds in developing algorithms to estimate the nonzero parameters. Of course, it is desirable to establish the statistical consistency of the estimates under a new asymptotic regime where both $n, p \to \infty$. Interestingly, it has emerged from such asymptotic theory that for the consistency to hold in some generic problems, the dimensions n, p of the data and the *sparsity index* of the model must satisfy

$$\frac{\log p}{n} \cdot s(\beta) \ll 1. \tag{1.1}$$

The ratio $\log p/n$ does play a central role in establishing consistency results for variety of covariance estimators proposed for high-dimensional data in recent years (Bühlmann and van de Geer, 2011).

1.1 LEAST SQUARES AND REGULARIZED REGRESSION

The idea of least squares estimation of the regression parameters in the familiar linear model

$$\boldsymbol{Y} = \boldsymbol{X}\beta + \boldsymbol{\varepsilon}, \tag{1.2}$$

has served statistics quite well when the sample size n is large and p is *fixed* and small, say less than 50. The principle of model simplicity or *parsimony* coupled

with the techniques of subset, forward, and backward selections have been developed and used to fit such models either for the purpose of describing the data or for its prediction.

The whole machinery of least-squares fails or does not work well for the high-dimensional data where the ubiquitous $\boldsymbol{X}'\boldsymbol{X}$ matrix is not invertible. The traditional remedy is the ridge regression (Hoerl and Kennard, 1970), which replaces the residual sum of squares of errors by its penalized version:

$$Q(\beta) = \|\boldsymbol{Y} - \boldsymbol{X}\beta\|^2 + \lambda \sum_j |\beta_j|^2, \tag{1.3}$$

where $\lambda > 0$ is a penalty controlling the length of the vector of regression parameters. The unique ridge solution is

$$\widehat{\beta}_{\text{ridge}} = (\boldsymbol{X}'\boldsymbol{X} + \lambda \mathbf{I})^{-1}\boldsymbol{X}'\boldsymbol{Y}, \tag{1.4}$$

which amounts to adding λ to the diagonal entries of $\boldsymbol{X}'\boldsymbol{X}$, and then inverting it. The ridge solution works rather well in the presence of multicollinearity and when p is not too large; however, in general it does not induce sparsity in the model. Nevertheless, it points to the fruitful direction of penalizing a norm of the high-dimensional vector of coefficients (parameters) in the model.

In the modern context of high-dimensional data, the standard goals of regression analysis have also shifted toward:

(I) construction of *good predictors* where the actual values of coefficients in the model are *irrelevant*;

(II) giving causal interpretations of the factors and determining which variables are more *important*.

It turns out that regularization is important for both of these goals, but the appropriate magnitude of the regularization parameter depends on which goal is more important for a given problem. Historically, Goal (II) has been the engine of the statistical developments and the thought of irrelevancy of the parameter values was not imaginable. However, nowadays Goal (I) is the primary focus of developing algorithms in the machine learning theory (Bühlmann and van de Geer, 2011). In general, the pair $(\boldsymbol{Y}, \boldsymbol{X})$ is usually modeled nonparametrically as

$$\boldsymbol{Y} = m(\boldsymbol{X}) + \boldsymbol{\varepsilon}, \tag{1.5}$$

with $E(\boldsymbol{\varepsilon}) = 0$ and $m(\cdot)$ a smooth unknown function. Then using a family of basis functions $b_j(\boldsymbol{X})$, $j = 1, 2, \cdots$, $m(\cdot)$ is approximated closely with the sums: $\sum_{j=1}^{p} \beta_j b_j(\boldsymbol{X})$, for a large p, where $\beta = (\beta_1, \cdots, \beta_p)'$ is a vector of coefficients. Of course, when $b_i(\boldsymbol{X}) = \boldsymbol{X}_j$ is the jth column of the design matrix, then the approximation scheme reduces to the familiar linear regression model.

In the high-dimensional contexts, the ridge or ℓ_2 penalty on $||\beta||_2^2 = \sum_{j=1}^{p} \beta_j^2$ has lost some of its attractions to competitors like the Lasso which penalizes large values of the sum $\sum_{j=1}^{p} |\beta_j| = ||\beta||_1$ and hence forces many of the smaller β_j's to be estimated by zero. This zeroing of the coefficients is seen as *model selection* in the sense that only variables with $\beta_j \neq 0$ are included. To this end, perhaps a more natural penalty function is

$$P(\beta) = \sum_{j=1}^{p} I(\beta_j \neq 0) = ||\beta||_0, \tag{1.6}$$

which counts the number of nonzero coefficients. However, the ℓ_0 norm is not an easy function to work with so far as optimization is concerned as it is neither smooth nor convex. Fortunately, the Lasso penalty is the closest convex member of the family of penalty functions of the form $||\beta||_\alpha^\alpha = \sum_{j=1}^{p} |\beta|^\alpha$, $\alpha > 0$ to (1.6).

1.2 LASSO: SURVIVAL OF THE BIGGER

In this section, we indicate that the Lasso regression which corresponds to replacing the ridge penalty in (1.3) by the ℓ_1 penalty on the coefficients leads to more sparse solutions than the ridge penalty. It forces to zero the smaller coefficients, but keeps the bigger ones around.

The Lasso regression is one of the most popular approaches for selecting significant variables and estimating regression coefficients simultaneously. It corresponds to a penalized least-squares regression with the ℓ_1 penalty on the coefficients (Tibshirani, 1996). Compared to the ridge penalty, it minimizes the sum of squares of residuals subject to a constraint on the sum of absolute values of the regression coefficients:

$$Q(\beta) = \frac{1}{2}\|\boldsymbol{Y} - \boldsymbol{X}\beta\|^2 + \lambda \sum_j |\beta_j|, \tag{1.7}$$

where $\lambda > 0$ is a penalty or tuning parameter controlling the sparsity of the model or the magnitude of the estimates. It is evident that for larger values of the tuning parameter λ, the Lasso estimate $\widehat{\beta}(\lambda)$ obtained by minimizing (1.7) shrinks or forces the regression coefficients toward zero. Since the regularization parameter controls the model complexity, its proper selection is of critical importance in the applications of the Lasso and other penalized least-squares/likelihood methods. In most of what follows in the sequel, it is assumed that λ is fixed and known.

Unlike the closed-form ridge solution in (1.4), due to the nature of the constraint in (1.7), its solution is nonlinear in the responses Y_i's, see (1.11). Fundamental to understanding and computing the Lasso solution is the *soft-thresholding operator*.

Its relevance is motivated using the simple problem of minimizing the function (1.7) for $n = p = 1$, or for a generic observation y and $\boldsymbol{X} = 1$:

$$Q(\beta) = \frac{1}{2}(y - \beta)^2 + \lambda|\beta|, \tag{1.8}$$

for a fixed λ. Note that $|\beta|$ is a differentiable function at $\beta \neq 0$ and the derivative $Q(\cdot)$ with respect to such a β is

$$Q'(\beta) = -y + \beta + \lambda \cdot \text{sign}(\beta) = 0, \tag{1.9}$$

where $\text{sign}(\beta)$ is defined through $|\beta| = \beta \cdot \text{sign}(\beta)$. By convention, its value at zero is set to be zero. The explicit solution of β in terms of y, λ from (1.9) is

$$\widehat{\beta}(\lambda) = \text{sign}(y)(|y| - \lambda)_+, \tag{1.10}$$

where $(x)_+ = x$, if $x > 0$ and 0, otherwise (see Section 2.5 for details). This simple closed-form solution reveals the following two fundamental characteristics of the Lasso solution:

1. Increasing the penalty λ will shrink the Lasso estimate $\widehat{\beta}(\lambda)$ toward 0. In fact, as soon as λ exceeds $|y|$, the Lasso estimate $\widehat{\beta}(\lambda)$ becomes zero and will remain so thereafter
2. The Lasso solution is *piecewise linear* in the penalty λ (see Fig. 1.1).

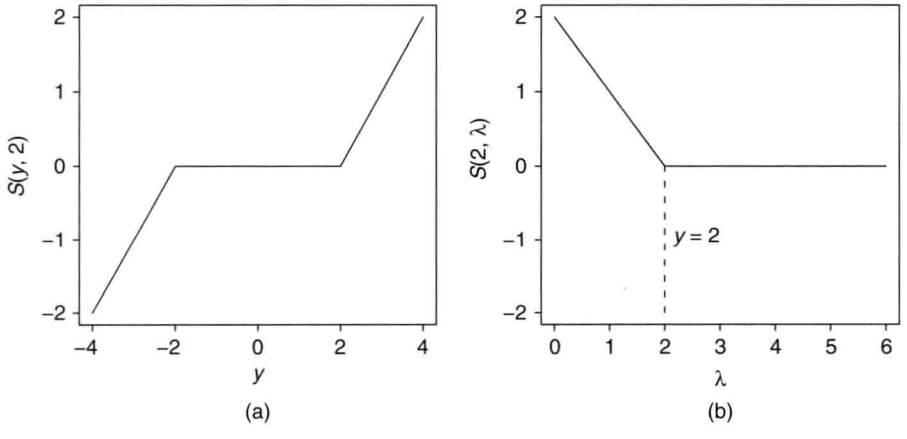


FIGURE 1.1 (a) Plot of the soft-thresholding operator for $\lambda = 2$. (b) Piecewise linearity of Lasso estimator for $y = 2$.